MORRIS D. FRIEDMAN- 572 California Street -Newtonville 60, Massachusetts, U.S.A.

Russian Translation

D / 128

# On Sufficient and Necessary Statistics for A Family of Probability Distributions



E. B. DYNKIN

Doklady, AN USSR, vol. 75, No. 2, 1950, pp. 161-164

Investigated in this work is the general problem of calculating sufficient statistics (see [1], pp. 530-531 and also [2]) for a given family of one-dimensional probability distributions. We will analyze statistics assuming a vector value; hence, obviously, the necessity of studying the sufficient system of statistics separately drops out. Definition 2 introduces a new concept of necessary statistics. If the knowledge of any sufficient statistics yields sufficient material to estimate unknown parameters which are sufficient in the sense that knowledge of the total result of observation does not add anything essential to this material, then the knowledge of the whole necessary statistics is necessary so that no essential loss of information would occur. A single necessary and sufficient statistic exists for each family of distributions with a fixed volume of choices, to the accuracy of equivalence. Theorem 1 gives a method of calculating it. Meanwhile, Theorem 1 introduces a new concept of the rank of a family of distributions. The concept of sufficient statistics for a family of infinite rank is fruitless. All families of finite rank are separated out by theorem 2 (a particular case of the latter is the Darmois theorem [3]). Families of distributions, obtained from certain distributions of linear transformations of a straight line, play an important part in mathematical statistics. Theorems 3 and 4 are devoted to a special investigation of such families of distributions.*

We will analyze the family $\mathfrak{G}$ of one-dimensional distributions $P_\theta(A)$ (the

* In particular, these theorems give an answer to the question, posed by A. N. Kolmogorov, on what shape the sufficient statistics, for the families of distributions mentioned, can have.

parameter $\Theta$ runs through a certain auxiliary manifold $S$). Hence, we assume that each $P_\Theta(A)$ distribution is given in a certain $\Delta$ interval (finite or infinite) by the density $p(x,\Theta)$ which is a positive piece-wise smoot function (regularity condition).* Let a series of $n$ independent experiments be made, where the results of each experiment are subject to the identical distribution law from the class $\mathfrak{S}$. The possible results generate an $n$-dimensional space $R^n$.

Definition 1. Each function $\chi(x_1,\ldots,x_n)$, assuming values from a certain vector space $R^m$ determined for all $x_1 \in \Delta, \ldots, x_n \in \Delta$ and satisfying the relation:

$$p(x_1,\Theta)p(x_2,\Theta)\ldots p(x_n,\Theta) = \bar{p}\Big(\chi(x_1,x_2,\ldots,x_n),\Theta\Big)\, q(x_1,x_2,\ldots,x_n)$$

in this region is called a SUFFICIENT STATISTIC FOR A FAMILY OF DISTRIBUTIONS IN AN INTERVAL OF $n$ VOLUME CHOICES.

Let us make the abbreviation $(x_1,x_2,\ldots,x_n) = x$. Let $\chi_1(x)$ and $\chi_2(x)$ be two functions defined in a certain region $G$ of the $R^n$ space. Conditionally, let us say that $\chi_2$ is subject to $\chi_1$ if $\chi_2(x') = \chi_2(x'')$ results from $\chi_1(x') = \chi_1(x'')$. We will call $\chi_1$ and $\chi_2$ equivalent if each is subject to the other. The $\varepsilon(x) = x$ statistic is sufficient for any distribution system. We will say that the $\chi(x)$ statistic is trivial in the region $G$ if it is equivalent to $\varepsilon(x)$ in a certain region $\tilde{G} \subset G$.

Definition 2. Each function defined for $x_1 \in \Delta, x_2 \in \Delta, \ldots, x_n \in \Delta$ and subject to any sufficient statistic in this region is called a NECESSARY STATISTIC FOR THE FAMILY OF DISTRIBUTIONS $\mathfrak{S}$ IN THE $\Delta$ INTERVAL OF $n$ VOLUME CHOICES.

THEOREM 1. Let $\Theta_o$ be an arbitrary element of $S$. Let us put $g_x(\Theta) = \ln p(x,\Theta) - \ln p(x,\Theta_o)$ and let us denote through $L$ the minimum linear space of functions, defined in $\Delta$, containing constants and containing $g_x(\Theta)$

* We call the function $p(x)$ piece-wise smooth in $\Delta$ if a region $G \subset \Delta$ exists such that $\bar{G} = \bar{\Delta}$ and $\frac{dp(x)}{dx}$ exists and is continuous in $G$.

for any $\Theta \in S$. Let the dimensionality of $L$ be $r+1$ (that $r = \infty$ is not excluded). Then:

A. For every finite $n \leq r$, the arbitrary sufficient statistic for the family $\mathfrak{S}$ in the $\Delta$ interval of $n$ volume choices is trivial.

B. If the functions $1, \varphi_1(x), \varphi_2(x), \ldots, \varphi_r(x)$ form a basis in $L$, then for any $n \geq r$, the system of functions

$$\gamma_i(x_1, x_2, \ldots, x_n) = \varphi_i(x_1) + \varphi_i(x_2) + \ldots + \varphi_i(x_n) \quad (i=1,2,\ldots,r;\ x_1, \ldots, x_n \in \Delta)$$

is functionally independent and forms a necessary and sufficient statistic for the family $\mathfrak{S}$ in the $\Delta$ interval of $n$ volume choices.

We call the number $r$, defined in THEOREM 1, THE RANK OF THE SYSTEM OF DISTRIBUTIONS $\mathfrak{S}$ IN THE $\Delta$ INTERVAL.

THEOREM 2. In order that the system of distributions $\mathfrak{S}$ have a finite rank in the $\Delta$ interval, it is necessary and sufficient that the density $p(x,\Theta)$ be represented thus:

$$p(x,\Theta) = \exp\left(\sum_{i=1}^{r} \varphi_i(x) c_i(\Theta) + c_o(0) + \varphi_o(x)\right) \qquad (x \in \Delta,\ \Theta \in S)$$

where $\varphi_1(x), \ldots, \varphi_r(x)$ are piece-wise smooth in the $\Delta$ interval. Hence, if $1, \varphi_1(x), \ldots, \varphi_r(x)$ is a linearly independent system of functions, then the rank of $\mathfrak{S}$ is $r$ and for $n \geq r$, the system of functions:

$$\gamma_i(x_1, \ldots, x_n) = \varphi_i(x_1) + \varphi_i(x_2) + \ldots + \varphi_i(x_n) \qquad (i=1,2,\ldots,r;\ x_1, \ldots, x_n \in \Delta)$$

is functionally independent and forms a necessary and sufficient statistic for $\mathfrak{S}$ in $n$ volume choices.

THEOREM 2. Let the $F(x,\Theta)$ distribution function be compared to each $\Theta \in S$. Let the family $\mathfrak{S}$ of these distributions satisfy the regularity condition in the $\Delta$ interval. Then:

A. If the $\mathfrak{S}_1$ family of $F(x-\alpha,\Theta)$ $(|\alpha| < \delta, \Theta \in S)$ distributions has a

finite rank in $\Delta$ for a certain $\delta > 0$, then the $p(x,\theta)$ density is represented as:

$$(1) \qquad p(x,\theta) = \exp\left(\sum_{i=1}^{s} c_i(\theta)\, x^{n_i} e^{\mu_i x}\right) \qquad (x \in \Delta,\ \theta \in S)$$

where $\mu_i$ are complex, $n_i$ are constant integers, $c_i(\theta)$ are functions taking on complex values.

B. If $\Delta$ does not contain zero and the $\mathfrak{S}_2$ family of distributions $F(\frac{x}{\sigma},\theta)$ $(\theta \in S, \frac{1}{p} < \sigma < p)$ has a finite rank in $\Delta$ for a certain $p > 1$, then the density $p(x,\theta)$ is represented thus:

$$(2) \qquad p(x,\theta) = \exp\left(\sum_{i=1}^{s} c_i(\theta)(\ln|x|)^{n_i}|x|^{\mu_i}\right) \quad (x \in \Delta,\ \theta \in S)$$

$[\mu_i,\ n_i,\ c_i(\theta)$ as in (1)$]$.

C. If the $\mathfrak{S}_3$ family of distributions $F\left(\frac{x-\alpha}{\sigma},\theta\right)$ $(\theta \in S, |\alpha| < \delta, \frac{1}{p} < \sigma < p)$ has finite rank in $\Delta$ for certain $\delta > 0$ and $p > 1$, then:

$$(3) \qquad p(x,\theta) = \exp Q(x,\theta) \qquad (x \in \Delta, \theta \in S)$$

where $Q(x,\theta)$ is a polynomial in $x$ with complex coefficients dependent on $\theta$.

A necessary and sufficient statistic for the $\mathfrak{S}_1$ family can be formed from functions of the form $x_1^k e^{\lambda x_1} + x_2^k e^{\lambda x_2} + \ldots + x_n^k e^{\lambda x_n}$; for the $\mathfrak{S}_2$ family, from functions of the form $|x_1|^\lambda (\ln|x_1|)^k + |x_2|^\lambda (\ln|x_2|)^k + \ldots + |x_n|^\lambda (\ln|x_n|)^k$ and for the $\mathfrak{S}_3$ family, from functions of the form $x_1^k + x_2^k + \ldots + x_n^k$ ($k$ is an integer, $\lambda$ is a complex constant).*

The probability density function for many important distributions equals zero outside a certain interval. If the whole line be considered as the $\Delta$ interval, then the regularity condition is not fulfilled here and THEOREM 3 is not applicable directly. The following theorem can be used instead.

---

* It is not difficult to indicate that set of $(k,\lambda)$ pairs to which corresponds the system of functions giving a necessary and sufficient statistic. We omit the formulation of the appropriate rule because of insufficient space.

THEOREM 4. Let the $p(x,\theta)$ function for each $\theta$ of $S$ be the probability density of a certain one-dimensional distribution. Let $p(x,\theta)$ be positive and piece-wise smooth with respect to $x$ in a certain $\Delta$ interval for all $\theta \in S$ and equal to zero outside $\Delta$. Let us denote through $\mathfrak{G}_1$ the family of distributions $p(x-\alpha,\theta)$ $(\theta \in S, -\infty < \alpha < +\infty)$, through $\mathfrak{G}_2$ the family of distributions $\frac{1}{\sigma} p(\frac{x}{\sigma},\theta)$ $(\theta \in S,\ 0 < \sigma < \infty)$ and through $\mathfrak{G}_3$ the family of distributions $\frac{1}{\sigma} p\left(\frac{x-\alpha}{\sigma},\theta\right)$ $(\theta \in S, -\infty < \alpha < +\infty,\ 0 < \sigma < +\infty)$. In order for the $\mathfrak{G}_1$, $\mathfrak{G}_2$, $\mathfrak{G}_3$ families to have finite rank on the whole line, it is necessary and sufficient that the $p(x,\theta)$ density be represented for $\theta \in S$, $x \in \Delta$, respectively, as (1), (2) or (3) (it is also assumed in the case of $\mathfrak{G}_2$ that the $\Delta$ interval does not contain zero).

A necessary and sufficient statistic for the $\mathfrak{G}_1$ and $\mathfrak{G}_2$ families is given by the system of functions mentioned in THEOREM 3 if $\Delta = (-\infty, +\infty)$. This system of functions must be supplemented by the function $\min(x_1,x_2,\ldots,x_n)$ in the $\Delta = (a,+\infty)$ case ($a$ finite) and by the function $\max(x_1,x_2,\ldots,x_n)$ in the $\Delta = (-\infty,b)$ case ($b$ finite). Finally, both the functions $\min(x_1,x_2,\ldots,x_n)$ and $\max(x_1,x_2,\ldots,x_n)$ must be added to the system of functions of THEOREM 3 in the $\Delta = (a,b)$ case, where $a$ and $b$ are finite. Similarly, the necessary and sufficient statistic for the $\mathfrak{G}_2$ family coincides with that mentioned in THEOREM 3 if $\Delta = (0,+\infty)$ or $\Delta = (-\infty,0)$; it is obtained from the statistic mentioned in THEOREM 3 by adding the $\min(x_1,x_2,\ldots,x_n)$ if $\Delta = (a,+\infty)$ $(a>0)$ or $\Delta = (-\infty,b)$ $(b<0)$; by adding the $\max(x_1,x_2,\ldots,x_n)$ if $\Delta = (0,a)$ $(0 < a < +\infty)$ or $\Delta = (b,0)$ $(-\infty < b < 0)$ and by adding both $\max(x_1,\ldots,x_n)$ and $\min(x_1,\ldots,x_n)$ if $\Delta = (a,b)$ where $a$ and $b$ are finite and not zero.

Examples.

1. The Gaussian density $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$ has the form (3). In

conformance with THEOREM 3 , the pair of functions $\left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)$ is a necessary and sufficient statistic for the family of $\frac{1}{\sigma} p\left(\frac{x-\alpha}{\sigma}\right)$ distributions.

2. Let us investigate the family of distributions given on the positive semi-axis by the density:

$$p(x,\beta,\gamma,\mu) = Cx^{\beta} e^{-\gamma x^{\mu}} \tag{4}$$

If $\mu$ is known, then the necessary and sufficient statistic is given by the pair of functions $\left(\sum_{i=1}^{n} x_i^{\mu}, \sum_{i=1}^{n} \ln x_i\right)$ for unknown $\beta$ and $\gamma$ ,* by the functions $\sum_{i=1}^{n} \ln x_i$ for unknown $\beta$ and known $\gamma$ and by the functions $\sum_{i=1}^{n} x_i^{\mu}$ for unknown $\gamma$ and known $\beta$ . If $\mu$ is unknown, then the rank of the family (4) is infinite in any $\Delta$ interval and this means that there are no non-trivial statistics (this was first proven by Pinsker).

3. For the $p(x-\alpha,\beta,\gamma,\mu)$ family, where $p(x,\beta,\gamma,\mu)$ is determined by (4), the necessary and sufficient statistic is obtained for known $\mu$ , according to THEOREM 4, by adding to the functions mentioned in example 2, the function $\min(x_1,x_2,\ldots,x_n)$ . The rank of the family considered is infinite for unknown $\mu$ .

4. The density:

$$p(x,\theta,\alpha_1,\sigma_1,\alpha_2,\sigma_2) = \frac{\theta}{\sqrt{2\pi}\,\sigma_1} \exp\left[-\frac{(x-\alpha_1)^2}{2\sigma_1^2}\right] + \frac{(1-\theta)}{\sqrt{2\pi}\,\sigma_2} \exp\left[-\frac{(x-\alpha_2)^2}{2\sigma_2^2}\right]$$

$(0 < \theta < 1)$ is obtained by mixing two Gaussian densities. If even one of the $\theta,\alpha_1,\sigma_1,\alpha_2,\sigma_2$ parameters is unknown, then the corresponding family of distributions has infinite rank.

5. The necessary and sufficient statistic for the family of Laplace distributions $p(x,\alpha,\sigma) = \frac{1}{2\sigma} \exp\left[-\frac{|x-\alpha|}{\sigma}\right]$ for known $\alpha$ equals $\sum_{i=1}^{n} |x_i - \alpha|$ . The family

---

* The expression '$\mu$ known, $\beta$, $\gamma$ unknown' means that the family of $p(x,\beta,\gamma,\mu)$ distributions is analyzed, where $\mu$ is fixed and $\beta$ and $\gamma$ vary in certain intervals.

has rank $\infty$ for unknown $\alpha$.

6. The density $p(x) = e^{-(x + e^{-x})}$ is encountered when investigating the limiting behavior of the maximum of $m$ independent random quantities as $m \to \infty$. The necessary and sufficient statistic is $\sum_{i=1}^{n} \exp(-x_i)$ (see THEOREM 3) for the family of $p(x-\alpha)$ distributions.

7. Let us consider the arbitrary $(a,b)$ interval and the uniform distribution in this interval. The family of all such distributions is obtained from any one of them by a linear transformation. The necessary and sufficient statistic is given by the pair of functions $\min(x_1,x_2,\ldots,x_n)$ and $\max(x_1,x_2,\ldots,x_n)$ (this example was first considered by A. N. Kolmogorov [4]).

June, 1950

References

1. H. CRAMER: Math. Methods of Statistics, 1946

2. A. N. KOLMOGOROV: Izv. AN USSR, math. ser., 14, No. 4, 1950

3. G. DARMOIS: C. R., 200, 1265, 1935

4. A. N. KOLMOGOROV: Izv. AN USSR, math. ser., 1942